# Performance Transfer and Behavioral Reliance in AI-Assisted Cybersecurity Training

Kameron Clark
Computer Information Systems
Georgia State University
Atlanta, United States of America
kclark94@gsu.edu

Benjamin M. Ampel
Computer Information Systems
Georgia State University
Atlanta, United States of America
bampel@gsu.edu

Balasubramaniam Ramesh
Computer Information Systems
Georgia State University
Atlanta, United States of America
bramesh@gsu.edu

***Abstract*—Generative AI can enhance cybersecurity training performance with support, but better-supported performance does not necessarily mean durable unaided skills. This six-week pilot study explores the performance-transfer gap in AI-assisted cybersecurity training, analyzing data from graduate students who completed cybersecurity detection tasks under control and AI-supported conditions. There were 167 task rows: 112 with AI support and 55 without. Participants using AI showed higher aided task accuracy than both their pretest and posttest unaided accuracy. However, unaided posttest accuracy did not significantly improve over pretest scores. These results indicate weak transfer or skill plateau, not erosion. No significant differences were found across conditions in learning gain, posttest accuracy, or transfer gap. Baseline skill influenced learning gains, with participants with low baseline skill improving more. Behavioral analyses revealed reliance as a significant negative predictor and perceived competence as a significant positive predictor of learning gain, though high collinearity among the behavioral measures means these indicators are best interpreted collectively. The study emphasizes distinguishing supported task performance from long-term skills and suggests reliance as a warning sign rather than a cause of skill erosion.**

**Keywords—Generative AI, cybersecurity training, performance transfer, AI reliance, cognitive offloading**

## I. Introduction

Organizational cybersecurity programs commonly deploy AI-supported decision aids to assist users with critical tasks such as phishing detection [2]. Such aids are typically evaluated by whether they improve task performance [1]. But cybersecurity evaluation must also ask whether that performance becomes skill independent of AI assistance, since detection work requires judgment amid ambiguity and changing threat conditions [5], [7]. A learner may correctly classify a phishing email when an AI recommendation is available but not internalize the cue-based reasoning needed to classify a similar or novel case without the tool.

Prior work on human-AI collaboration shows that AI support can improve performance when it helps humans and AI contribute complementary strengths [1]. Automation research also shows that reduced manual engagement can weaken situation awareness and make users less prepared to intervene when automation fails [3], [4]. Cognitive offloading research explains that people move memory, monitoring, or reasoning work into external tools when those tools reduce cognitive effort [5], [6]. AI support can improve task completion while changing what users practice [1], [5].

Cybersecurity makes the problem non-trivial as it is dynamic, adversarial, and nonstationary [7], [9]. Attackers frequently adapt tactics, alter cue structures, and exploit blind spots in established routines [7], [9]. Therefore, skill-degradation solutions from other domains may not transfer to cybersecurity. A fixed refresher task or a static checklist may preserve performance on known cases but may not preserve adaptive threat reasoning under adversarial novelty.

To help address this issue, we ran a pilot study to determine how repeated reliance on generative AI (GenAI) assistance influences the acquisition, calibration, and possible erosion of cybersecurity skills over time. We frame this gap as a performance-transfer problem in AI-assisted cybersecurity training. This pilot asks whether AI-supported task gains transfer to unaided cybersecurity detection skill, whether condition-level differences appear across control, low AI assistance, and high AI assistance, and whether behavioral indicators such as trust, reliance, perceived competence, collaboration, and cognitive effort help explain weaker learning gains. The concern is that task performance is not durable once support is removed. The pilot suggests a supported-performance advantage but no unaided pre-post improvement.

This paper makes four contributions. First, the pilot shows that participants perform better with AI support but do not clearly improve after the support is removed. Second, it reframes the outcome as weak transfer or skill plateau rather than skill erosion, since the pilot did not show reliable unaided decline. The key point is that supported performance did not translate into durable unaided skills. Third, it cautions against overinterpreting condition-level effects, as no significant differences were found between control, low AI, and high AI conditions in learning or transfer. Baseline skill influenced gains, with participants with low baseline skill improving more. Fourth, it identifies behavioral reliance as a potential warning sign, with reliance negatively predicting learning gain.

## II. Background and Literature Review

### A. Cybersecurity as a Boundary Condition

Cybersecurity is an adversarial domain in which attackers adapt their behavior in response to defenses [7], [9]. Veksler et al. emphasize the role of human cognition, expert judgment,

and attacker-defender behavior in cybersecurity tasks [9]. Recent work on GenAI and zero-trust erosion also shows how AI-enabled attacks can mutate identity cues, telemetry, and trust signals [11]. This makes cybersecurity a boundary condition for standard skill-degradation logic. In more stable domains, skill maintenance can rely more heavily on recurring practice against known task structures. In cybersecurity, the skill set must adapt: the learner needs not just recognition of familiar phishing cues but adaptive skepticism when a case does not fit prior patterns. A plateau in unaided skill is therefore not merely a weak learning outcome; it may be an early warning sign. If this domain raises the stakes for skill maintenance, the next question is how reliance on automated aids can quietly undermine it.

### *B. Automation Reliance and Out-of-the-loop Risk*

Automation can improve performance while weakening active engagement [3], [4]. Endsley and Kiris describe the out-of-the-loop performance problem, in which operators in automated systems may lose the situational awareness needed to take over when automation fails [4]. Parasuraman and Riley similarly distinguish productive use from misuse, disuse, and abuse of automation [6]. These concepts are useful for cybersecurity training because trainees may become monitors of AI output rather than active analysts of evidence. The cybersecurity translation is direct but not identical. The risk is not only that users accept a wrong recommendation. The deeper training risk is that users may stop practicing the component skills that make detection transferable. These skills include inspecting cues, weighing ambiguous evidence, noticing novelty, articulating a rationale, and deciding when to question a recommendation. Why such skills atrophy under reliance is clarified by the cognitive mechanism of offloading.

### *C. Cognitive Offloading and Learning Transfer*

Cognitive offloading refers to shifting cognitive work onto external resources to reduce mental demand [5]. It can be useful, but can also change what users internalize. Sparrow, Liu, and Wegner show that people may encode where to find information rather than the information itself when they expect future access to external systems [8]. In cybersecurity training, this means users may learn how to consult the tool rather than how to reason through the threat. The learning concern is not ordinary tool use but repeated substitution of the reasoning process. Durable skill acquisition requires active reasoning, practice, reflection, and feedback. If an AI repeatedly identifies the cues, explains the evidence, and frames the conclusion, the learner may receive correct answers without sufficient practice in developing independent judgment in detection. This risk is sharpened when the AI also explains its answers, because fluent explanations can shape how learners judge their own competence. Recent AI education research similarly warns that AI may enhance learning access while encouraging cognitive dependency if it replaces active recall or problem solving [5].

### *D. Human-AI Explanation, Calibration, and Deference*

Explainable AI does not automatically create better human-AI judgment. Bansal et al. found that explanations can increase acceptance of AI recommendations, including when the recommendation is wrong [1]. This matters for training because explanations may increase compliance without increasing independent understanding. Recent IS research also shows that AI explanations can shape metacognitive processes [10]. Metacognition concerns how people judge their own thinking and decide when to retain control or delegate to AI. In this paper, calibration is central. A learner may feel more capable after reading a fluent explanation, even if that explanation has not improved the learner's unaided detection skill. This paper treats behavioral deference as a possible mechanism rather than a confirmed effect. Reliance is especially important because it captures dependence on AI support, which may reduce independent cue evaluation. However, positive perceptions of AI are not automatically harmful. In the pilot results, perceived competence was positively associated with learning gain, which suggests that the behavioral mechanism is more nuanced than a simple deference story.

### *E. Synthesis and Research Questions*

The literature review suggests that cybersecurity raises the stakes for durable, transferable skills; reliance on automation and cognitive offloading describes how that skill can quietly erode when a capable aid is always available; and explanation and calibration describe how learners may misjudge their own competence. AI assistance may lift performance while support is present without producing learning that persists once it is removed, and behavioral reliance may serve as an observable marker of that gap. This study therefore poses the following question:

> *Does AI-assisted cybersecurity training improve unaided detection skills, and do behavioral indicators, such as reliance, help explain who benefits?*

We organize the investigation around four hypotheses.

**H1.** Participants will perform better on AI-supported tasks than on their unaided pretest and posttest tasks, but unaided posttest performance will not improve significantly over pretest performance.

**H2.** Learning gain, posttest accuracy, and the transfer gap will not differ significantly across the conditions.

**H3.** Baseline skill will shape learning gain, with lower-baseline improving more than higher-baseline participants.

**H4.** Behavioral reliance will be negatively associated with learning gain. Perceived competence will be positively associated with learning gain.

## III. Methodology

This pilot study examined the performance of AI-assisted cybersecurity detection among graduate-level information systems and cybersecurity students. Participants completed simulated cybersecurity detection tasks across a six-week study period. The tasks required participants to evaluate security-

related scenarios, including phishing-style detection judgments, and determine whether the material represented a potential threat.

### A. Study Context and Participants

The pilot study included 37 participants. The cleaned analytic dataset used for the reported analyses contained 167 complete task rows. Of these, 112 rows came from AI-supported conditions and 55 rows came from the control condition. Because the broader study involved repeated participation across weeks, participant identification was complicated by inconsistent ID entry and attrition.

### B. Study Design and Procedure

Participants completed the study through a structured survey-based task environment. Each task session followed the same general sequence: an unaided pretest, a middle task block, an unaided posttest, and post-task behavioral survey items. The pretest measured baseline unaided cybersecurity detection performance prior to the introduction of AI support. The middle task block measured task performance during the main experimental condition. The posttest measured unaided cybersecurity detection performance after the middle task block.

Participants were assigned to one of three conditions: Control, Low AI Assistance, or High AI Assistance. Participants in the control condition completed the middle task block without AI support. Participants in the AI-supported conditions received AI assistance during the middle task block. Because the control group did not receive AI support, the middle task block is described as task-block accuracy in full-sample analyses. It is described as aided accuracy only in analyses restricted to participants in the Low AI Assistance and High AI Assistance conditions. After completing the task blocks, participants answered behavioral survey items measuring trust, reliance, collaboration, perceived competence, and cognitive effort. These items were used to examine whether behavioral indicators helped explain variation in learning gain and transfer outcomes. We show this flow in Figure 1.

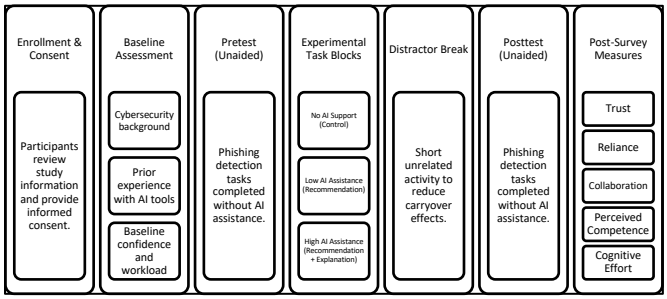

Fig. 1. Study Procedure Across The Six-Week Period

### C. Measures

Performance was measured using accuracy scores for the pretest, middle task block, and posttest. Pretest accuracy represented baseline unaided detection performance. Task-block accuracy represented performance during the middle task block. Aided accuracy referred only to the middle task block for participants who received AI support. Posttest accuracy represented unaided performance after the middle task block. We detail these measures in Table I.

TABLE I. OPERATIONAL DEFINITIONS FOR PILOT OUTCOMES

| Construct | Definition | Interpretation |
|---|---|---|
| Pretest Accuracy | Accuracy during the unaided pretest block. | Baseline unaided detection performance. |
| Task-Block Accuracy | Accuracy during the middle task block for the full sample. | Middle-block performance; not "aided" for control participants. |
| Aided Accuracy | Accuracy during the middle task block only for AI-supported participants. | Supported performance while AI is available. |
| Posttest Accuracy | Accuracy during the unaided posttest block. | Unaided performance after the middle task block. |
| Learning Gain | Posttest Accuracy − Pretest Accuracy. | Change in unaided detection performance. |
| Task-Block Advantage | Task-Block Accuracy − Pretest Accuracy. | Immediate task-block performance advantage. |
| Transfer Gap | Task-Block Accuracy − Posttest Accuracy. | Gap between middle-block performance and later unaided performance. |
| Skill Plateau | No statistically reliable improvement from pretest to posttest. | Weak transfer; not proof of erosion. |

The behavioral constructs were measured using post-task survey items. Trust, collaboration, perceived competence, and cognitive effort were measured using three-item scales. Reliance was included as a behavioral indicator of dependence on AI support. Internal consistency was strong for the multi-item scales, with Cronbach's alpha values above .94 for trust, collaboration, perceived competence, and cognitive effort. We show the results in Table II.

TABLE II. BEHAVIORAL CONSTRUCTS IN THE PILOT TRAINING

| Construct | Working Meaning | Use in Analysis |
|---|---|---|
| **Trust** | Confidence that the AI tool is dependable. | Measured with a three-item scale; alpha = .957. |
| **Reliance** | Self-reported on AI support. | Single construct as negative predictor of gain. |
| **Collaboration** | Perception that the participant and AI worked together. | Measured with a three-item scale; alpha = .962. |
| **Perceived Competence** | Belief that the tool improved ability or task effectiveness. | Measured with a three-item scale; alpha = .947. |
| **Cognitive Effort** | Self-reported effort invested in the task. | Measured with a three-item scale; alpha = .951. |

### D. Score Normalization and Data Preparation

The middle task block generally contained more items than either the pretest or posttest block individually. To make scores comparable across blocks, all performance scores were converted into accuracy proportions on a 0–1 scale. This normalization allowed pretest, task-block, and posttest performance to be compared despite differences in item counts. However, the middle-block score may be more stable because it was based on more items. For that reason, the strongest transfer comparison is the unaided posttest versus unaided pretest comparison.

Rows with incomplete performance data were removed from the main paired analyses. The final analytic dataset contained 167 complete task rows, including 112 AI-supported rows and 55 control rows.

### *E. Analysis Strategy*

The analysis followed the paper's performance-transfer logic. First, paired tests among AI-supported participants compared aided-task accuracy with unaided pretest and posttest accuracy to see if performance was higher with AI support. Second, unaided posttest accuracy was compared with pretest to assess transfer of supported performance into unaided skill. Third, ANOVA tests were conducted to assess differences among the control, low, and high AI Assistance groups. ANCOVA models tested condition effects, controlling for pretest accuracy and week. Additional analyses examined baseline performance, behavioral predictors, and manipulation checks. Baseline-group analyses explored learning gains across different baseline levels. Behavioral analyses assessed if trust, reliance, collaboration, perceived competence, and cognitive effort explained learning gains, though results should be viewed cautiously due to the exploratory nature and overlapping predictors.

## IV. Findings

The findings are organized around the paper's central performance-transfer question. First, we examine whether AI-supported participants performed better while support was available and whether that advantage transferred to unaided posttest performance. Second, we test whether performance differed across control, low AI Assistance, and high AI Assistance conditions. Third, we examine whether baseline skill shaped learning gains. Finally, we evaluate whether behavioral indicators help explain variation in learning gain. This structure separates the main empirical pattern from supporting analyses and avoids treating every statistical test as equally central.

### *A. Supported Performance Improved, Unaided Transfer Weak*

To test whether the supported task-block advantage carried over to unaided skill, we conducted a set of paired comparisons of detection accuracy within the AI-supported group. Table III reports these tests, contrasting aided accuracy against both unaided pretest and posttest accuracy and comparing unaided posttest against pretest accuracy.

TABLE III. Main Paired Test

| Comparison | n | Mean 1 | Mean 2 | Difference | p |
|---|---|---|---|---|---|
| Aided vs. Pretest | 112 | .807 | .728 | +.080 | .012 |
| Aided vs. Posttest | 112 | .807 | .699 | +.108 | .001 |
| Posttest vs. Pretest | 112 | .699 | .728 | -.028 | .388 |

Among participants who received AI support, aided-task accuracy was significantly higher than unaided pretest accuracy, $t(111) = 2.55$, $p = .012$, with a mean difference of .080, 95% *CI* [.018, .142], Cohen's $dz = .24$. Aided-task accuracy was also significantly higher than unaided posttest accuracy, $t(111) = 3.31$, $p = .001$, with a mean difference of .108, 95% *CI* [.043, .172], Cohen's $dz = .31$. These results show a supported task-block advantage among AI-supported participants.

However, unaided posttest did not significantly exceed unaided pretest accuracy among AI-supported participants, $t(111) = -0.87$, $p = .388$, with a mean difference of -.028 (95% *CI*: [-0.093, 0.036]). The Wilcoxon signed-rank test showed the same pattern, $p = .458$. This means the supported-performance advantage did not clearly transfer into stronger unaided posttest performance. The finding should be interpreted as weak transfer or skill plateau, not as proof of skill erosion.

A TOST equivalence check using +/- .05 accuracy bounds did not establish statistical equivalence between pretest and posttest performance. Therefore, the correct interpretation is not that pretest and posttest were proven equivalent. The correct interpretation is that the pilot failed to detect significant unaided improvement. These results support the main performance-transfer argument. AI-supported participants performed better while support was available, but that advantage did not clearly become durable unaided cybersecurity detection skill. This pattern supports H1.

### *B. Condition Assignment Did Not Explain Differences*

Condition-level tests showed no statistically significant differences across control, low AI Assistance, and high AI assistance for pretest accuracy, task-block accuracy, posttest accuracy, learning gain, task-block advantage, or transfer gap. This means the pilot does not support claims that low AI or high AI led to stronger, weaker, or eroded learning relative to the control. ANCOVA models led to the same cautious conclusion. In a model predicting middle task-block accuracy while controlling for pretest accuracy and week, the control group did not differ significantly from the AI-supported group ($p = .718$). In a model predicting posttest accuracy while controlling for pretest accuracy and week, neither high AI assistance ($p = .404$) nor low AI assistance ($p = .770$) significantly differed from the control. Table IV summarizes the condition-level tests across all outcomes.

TABLE IV. Conditional-Level Tests

| Outcome | p | Interpretation |
|---|---|---|
| Pretest accuracy | .906 | Groups did not differ at baseline. |
| Task-block accuracy | .435 | No significant condition difference during the middle block. |
| Posttest accuracy | .488 | No significant condition difference after support was removed. |
| Learning gain | .410 | No condition produced stronger unaided learning gains. |
| Task-block advantage | .528 | No condition-level difference in middle-block improvement. |
| Transfer gap | .952 | No condition-level difference in supported-to-unaided gap. |

These results limit the causal interpretation of the pilot. The strongest finding is not that one condition outperformed another. The stronger finding is that supported performance and unaided transfer behaved differently. The absence of condition-level differences supports H2.

### C. Baseline Skill Shaped Learning Gains

Baseline performance was an important factor in learning-gain patterns. Baseline groups were based on pretest accuracy, rather than experimental condition. Low-baseline participants improved from pretest to posttest, while middle- and high-baseline participants showed lower gains. This suggests that ceiling effects and regression to the mean partly shaped the flat overall transfer pattern. Table V reports learning gains for each baseline group.

TABLE V. LEARNING GAINS BY BASELINE PERFORMANCE GROUP

| Group | n | Pretest | Posttest | Learning Gain |
|---|---|---|---|---|
| Low baseline | 91 | .555 | .648 | +.093 |
| Middle baseline | 30 | .833 | .706 | -.128 |
| High baseline | 46 | 1.000 | .804 | -.196 |

Compared with the low-baseline group, the middle-baseline group showed significantly lower learning gain, $\beta = -.221$, $p < .001$, and the high-baseline group also showed significantly lower learning gain, $\beta = -.289$, $p < .001$. This does not eliminate the transfer-gap finding, but it shows that baseline ability must be considered when interpreting learning gains. Participants who started lower had more room to improve, while those who started higher had less room to improve further. This baseline-dependent pattern supports H3.

### D. Behavioral Evidence Was Mixed, but Reliance Stood Out

The behavioral measures showed high internal consistency with Cronbach's alpha over .951 for trust, collaboration, perceived competence, and cognitive effort. These support the use of combined scores, though some items may be redundant. Correlations between indicators and learning outcomes were mostly nonsignificant after correction, so mechanisms are not confirmed. Regression analyses indicated that adding behavioral indicators increased the explained variance from 29.8% to 40.9%, with reliance negatively predicting learning gain and perceived competence positively predicting it. Trust was marginal, and collaboration and effort were not significant. High collinearity among measures suggests they act together, with reliance as the clearest warning signal. Results support H4, but should be interpreted with caution due to collinearity.

### E. Conditions Shifted Subjective Experience More Than Performance

Manipulation checks showed that condition predicted several behavioral indicators even though it did not significantly predict the main performance outcomes. Trust, reliance, collaboration, and cognitive effort differed by condition, while perceived competence did not. Low AI Assistance was associated with lower trust and reliance relative to control, while High AI Assistance was associated with higher collaboration and higher cognitive effort. This suggests that the assistance manipulation affected participants' subjective experience more clearly than it affected measured learning outcomes.

## V. DISCUSSION

The pilot results suggest a performance-transfer gap: AI-supported participants performed better on supported tasks, but this did not clearly translate to improved unaided posttest performance. This indicates a weak transfer or a skill plateau, not erosion, and requires more long-term data. The results limit causal claims: no significant performance differences between groups were found, so AI did not cause skill loss or gain. Supported performance improved, but unaided transfer remained weak. Behavioral data show that reliance is linked to less learning, that perceived competence boosts learning, that trust is marginal, and that collaboration and effort are not significant predictors. This suggests that dependence on AI, not just positive perceptions, influences outcomes. In cybersecurity, this gap is critical. Training that rewards supported success without ensuring unaided transfer risks creating users who are effective with tools but less effective without them.

## VI. IMPLICATIONS FOR RESEARCH AND PRACTICE

### A. Implications for Research

Consistent with H1, AI-supported accuracy exceeded unaided performance while support was present, yet unaided posttest accuracy did not improve over pretest. Researchers should therefore treat supported performance and unaided transfer as distinct outcomes and measure what users can do once support is withdrawn, framing the gap as weak transfer or plateau rather than erosion.

Supporting H3, low-baseline participants improved while middle- and high-baseline participants showed lower gains, a pattern condition assignment did not explain. Future research should model baseline ability explicitly and build baseline stratification and item difficulty into sampling and analysis, since designs that ignore this moderation can mistake a ceiling effect or regression to the mean for an intervention effect.

Consistent with H4, reliance negatively predicted learning gain and perceived competence positively predicted it, but the high variance inflation among trust, reliance, and collaboration means these constructs were not cleanly separable here. Future studies should treat this result as exploratory and develop instruments that distinguish dependence, confidence, collaboration quality, and perceived learning sharply enough to isolate which mechanism drives the transfer gap.

Finally, because cybersecurity threats adapt, future designs should build adversarial novelty into the transfer measure. Separating familiar from novel threat cases would test whether the weak transfer seen here reflects a failure to internalize fixed cues or a deeper inability to reason adaptively when a case does not fit prior patterns.

### B. Implications for Practice

The study's central finding for practice is that supported accuracy overstates learning: consistent with H1, participants performed well while AI support was present but did not improve once it was removed. Training programs should therefore not certify AI-assisted learning on supported task accuracy alone and should build in unaided posttests and delayed transfer checks that measure what a learner retains

without the tool. A useful safeguard is to require learners to explain threat cues before any AI explanation is shown, so that genuine understanding is assessed rather than agreement with the system.

While instructors and organizations look for warning signs, this study points to reliance specifically, rather than to positive attitudes toward AI in general. Supporting H4, reliance was the clearest negative predictor of learning gain, whereas perceived competence predicted it positively; high confidence in the tool is not in itself a red flag. Monitoring should track behavioral dependence, such as how readily a learner defers to the system, rather than treating enthusiasm for AI as a problem to be managed.

Because the transfer gap is consistent with learners offloading the reasoning rather than acquiring it, AI training tools should be designed to keep that reasoning active. Tools should require learners to justify a decision, identify the cues behind it, and compare their reasoning against the AI's before accepting a recommendation. Feedback should grade the quality of that reasoning, not only the final answer, so that the practice that makes detection skill transferable is exercised even when a capable aid is available.

## VII. Limitations

This exploratory pilot involved 167 task rows. Participant-ID issues, attrition, and cross-group movement limit the validity of longitudinal conclusions. The middle task block had more items, and accuracy scores were normalized to 0-1, with the middle block likely more stable due to more observations. Future studies should use balanced blocks or item-level models. Condition effects were not significant, so no claims about AI impact on learning can be made. Behavioral analyses were exploratory; reliance negatively predicted learning gain, but confidence, reliance, and collaboration were highly correlated, so individual results should be interpreted with caution. The study doesn't directly test adversarial novelty but highlights its importance for transfer in cybersecurity. Future work should compare familiar and novel threats.

## VIII. Conclusion

We found evidence of a performance-transfer gap. AI-supported participants performed better during the supported task block, but unaided posttest accuracy did not significantly exceed that on the unaided pretest. Condition-level performance differences were not detected. Baseline ability shaped learning gains, and behavioral indicators added explanatory value, with reliance emerging as the clearest negative predictor.

The main contribution is a reframing of AI-assisted cybersecurity training evaluation. The question is not only whether AI improves performance while it is present, but also whether learners become better unaided analysts after support is removed. This pilot suggests that supported task success can coexist with weak unaided transfer. For cybersecurity training, that distinction matters because adaptive threat reasoning cannot be assumed from aided scores alone.


## Acknowledgment

This material is based upon work supported by the U.S. National Science Foundation under Award No. 2146497 through the NSF CyberCorps® Scholarship for Service (SFS) program at Georgia State University and by the U.S. Army Research Laboratory (ARL) under Cooperative Agreement Number W911NF-23-2-0224. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government.

AI Tools Used—AI tools were used to assist with language editing. The authors reviewed, revised, and approved all content and remain responsible for the originality, accuracy, and integrity of the manuscript.


## References


[1] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. S. Weld, "Does the whole exceed its parts? The effect of AI explanations on complementary team performance," in Proc. 2021 CHI Conf. Human Factors in Computing Systems, 2021, pp. 1–16. doi: 10.1145/3411764.3445717

[2] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," Telecommunication Systems, vol. 76, no. 1, pp. 139–154, Jan. 2021. doi: 10.1007/s11235-020-00733-2

[3] G. Chirayath, K. Premamalini, and J. Joseph, "Cognitive offloading or cognitive overload? How AI alters the mental architecture of coping," Frontiers in Psychology, vol. 16, Art. no. 1699320, 2025. doi: 10.3389/fpsyg.2025.1699320

[4] M. R. Endsley and E. O. Kiris, "The out-of-the-loop performance problem and level of control in automation," *Human Factors, vol. 37*, no. 2, pp. 381–394, 1995. doi: 10.1518/001872095779064555

[5] B. Jose, J. Cherian, A. M. Verghis, S. M. Varghise, M. S., and S. Joseph, "The cognitive paradox of AI in education: Between enhancement and erosion," Frontiers in Psychology, vol. 16, Art. no. 1550621, 2025. doi: 10.3389/fpsyg.2025.1550621

[6] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors, vol. 39*, no. 2, pp. 230–253, 1997. doi: 10.1518/001872097778543886

[7] E. F. Risko and S. J. Gilbert, "Cognitive offloading," *Trends in Cognitive Sciences, vol. 20*, no. 9, pp. 676–688, 2016. doi: 10.1016/j.tics.2016.07.002

[8] B. Sparrow, J. Liu, and D. M. Wegner, "Google effects on memory: Cognitive consequences of having information at our fingertips," Science, vol. 333, no. 6043, pp. 776–778, 2011. doi: 10.1126/science.1207745

[9] V. D. Veksler, N. Buchler, C. G. LaFleur, M. S. Yu, C. Lebiere, and C. Gonzalez, "Cognitive models in cybersecurity: Learning from expert analysts and predicting attacker behavior," Frontiers in Psychology, vol. 11, Art. no. 1049, 2020. doi: 10.3389/fpsyg.2020.01049

[10] M. von Zahn, L. Liebich, E. Jussupow, O. Hinz, and K. Bauer, "Knowing (not) to know: Explainable artificial intelligence and human metacognition," Information Systems Research, 2025. doi: 10.1287/isre.2024.1431

[11] D. Xu, I. Gondal, X. Yi, T. Susnjak, P. Watters, and T. R. McIntosh, "The erosion of cybersecurity zero-trust principles through generative AI: A survey on the challenges and future directions," Journal of Cybersecurity and Privacy, vol. 5, Art. no. 87, 2025. doi: 10.3390/jcp5040087